Inform : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol.8 No.1 January 2023, P-ISSN:2502-3470, E-ISSN:2581-0367

Comparison of Stemming Test Results of Tala Algorithms with Nazief Adriani in Abstract Documents and National News

Natalinda Pamungkas[1*], Erika Devi Udayanti[2], Bonifacius Vicky Indriyono[3], Wildan Mahmud[4], Ery Mintorini[5], Arika Norma Wahyu Dorroty[6], Sanina Quamila Putri [7]

[1,2,3,4,6,7]*Information Systems, Universitas Dian Nuswantoro, Semarang, Indonesia*

[5]*Visual Communication Design, Universitas Dian Nuswantoro, Semarang, Indonesia*

[1]natalinda.pamungkas@dsn.dinus.ac.id (*)

[2,3,4,5][erikadevi, ery.mintorini, wildan.mahmud, bonifacius.vicky.indriyono]@dsn.dinus.ac.id

[6,7][norma.wahyu08, saninap01]@gmail.com

Received: 2022-12-13; Accepted: 2023-01-09; Published: 2023-01-28

Abstract— The existence of information is undeniably needed by many people. This statement describes the increasing importance of information and the corresponding increase in the need for access to relevant documents and literature. The contents of the information derived from these documents are then sorted to make their meaning more understandable. This sorting process is known as stemming. Stemming is a process that is widely applied in basic word searches. Separating meaningless words can make information clearer. It is necessary to pay attention to the appropriate stemming algorithm according to the language used. Many stemming algorithms can be used to perform this basic word search process. Some of them are the Tala and Nazief Adriani algorithms. The two algorithms have differences in their work processes. The Tala algorithm adopts a rule-based Porter algorithm, while the Nazief & Adriani algorithm works based on a dictionary. The two algorithms have their respective advantages in terms of accuracy and speed. Therefore, in this study, an analysis will be carried out by comparing the performance of the two algorithms in the Indonesian language text-stemming process. The trial process uses several different data sources to measure the speed and accuracy of each algorithm. Data sources used in this study included abstracts of student thesis reports or final assignments of 30 students and information from online news as many as 200. From the results of the tests that have been carried out, it can be concluded that the Tala stemming algorithm has a lower accuracy level than Nazief Adriani. The Tala algorithm only has an average accuracy of 65.29%, while Nazief Adriani has an accuracy of 78.47%. Regarding speed, the Tala algorithm has a better speed than Nazief Adriani at 32.19 seconds and Nazief & Adriani at 65.2 seconds.

Keywords— Stemming, Nazief Adriani Algorithm, Tala Algorithm, Abstract Documents, National News

I. INTRODUCTION

The need for information in the current technological era is needed by its users. This technology is used in searching text documents, especially on the internet, to get information. The problem that often arises in getting information is finding information that fits your needs. Appropriate information can be obtained by performing word splitting in text documents. One way of separating words is to get information using the stemming process. Stemming is one of the processes of transforming words in the text into root words or removing word affixes [1]. The stemming algorithm for one language will differ from another. The Nazief Adriani and Tala algorithms are the most used for stemming documents in Indonesian. Nazief Adriani's algorithm is a stemming algorithm whose working principle uses a dictionary, while the Tala algorithm adopts Porter's algorithm and has a rule-based working principle. The stemming process in Indonesian language texts has a variety of affixes that must be removed to get the root word of a word [2]. Stemming can also be used in the Indonesian language learning process regarding basic words [3]. Nopiyanti [4] researched the formation of stemming applications used to search for basic words following the Big Language Dictionary Indonesia (KBBI). By using Porter's algorithm in the process of stemming 20 documents, it can still be developed by conducting tests with more documents and news datasets. The research conducted [1] analyzed the stages of stemming using Porter's algorithm for Indonesian documents using documents with the .txt extension/format and incomplete dictionary datasets. From this research, it is still possible to develop tests if the documents used can have the .pdf / .docx extension to determine the accuracy to be obtained. With a more equipped dictionary, accurate test results can be developed. Utomo [5] researched using an algorithm on the corpus (abstract). This research is based on the rules contained in the Tala algorithm. This research still has a high error rate in the stemming process, so it can still be developed. The research that increased the ability to stem by adding two levels of morphology was carried out by [6] to obtain quite a high accuracy. In this study, only using ten documents, development can still be carried out by comparison if tests are carried out on more news and document datasets. From the research conducted [2], comparing the results of the Porter and Nazief & Adriani algorithms, they still use only documents for testing. There is a possibility that can be done in getting accuracy in testing from several samples of news datasets to be a test that can be considered.

Several previous studies that address the topic of stemming algorithm performance. This journal [3] describes the process

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

of creating stemming software to search for a set of basic words that match those stored in the Big Indonesian Language Dictionary (KBBI) from Indonesian text documents using a porter stemmer. The researcher [7] describes the stemmer process in determining the classification of book types using the Porter-Stemmer algorithm. The study [7] discusses the use of Porter's algorithm for automatically classifying book types based on basic word search rules. The results of the study indicate that the algorithm is effective in accurately classifying book types. However, the research [4] mentioned suggests that rule-based algorithms can have a high error rate, which can negatively impact the accuracy of the final results. The study's results concluded [8] that the search process obtained an average application response time calculation of 5.66 seconds, and the recall results for searching documents obtained were an average of 83%. The research [9] on the Stemming algorithm for different tenses to improve the Persian dictionary explained a rule-based stemming algorithm, not using a dictionary. This study concludes [9] that the algorithm used has accuracy for ordinary verbs in simple/past, continuous and perfect tense but is limited only to dictionaries with regular verbs and is limited to testing 50 words out of 465 words in the dictionary. The Adriani & Nazief algorithm and the Similarity Algorithm can be used to check the title and thesis abstraction [10] and whether a title with that theme has been submitted. They are stemming functions to collect title indexes and thesis abstractions as a database so they can be checked using a similarity algorithm. A study [11] related to text similarity detection concluded that implementing the Nazief Adriani stemming method in the Rabin-Karp algorithm greatly affects the percentage level of text similarity, making it easier to detect text similarity.

Based on previous research literature on the stemming process, this study will compare the results of the Tala stemming test with Nazief Adriani on abstract documents and national news as the source data for testing.

II. RESEARCH METHODOLOGY

Research methodology is a method that can be studied scientifically and used by researchers to obtain data to support research activities or for other purposes [12].

A. Research Methods

This research was conducted using the experimental method. According to [13], experimental research is a research method that describes causal relationships, so this method can be said to be a causal research method. In particular, the research methods carried out can be described in Figure 1, it can be explained the stages of the research are as follows:

1. *Literature Study*: The first stage involves collecting several references supporting writing research topics. These sources may include academic journals, books, conference proceedings, and other articles.

2. *Data Collection*: The next research phase is data collection. The data used to implement this research comes from the essence of student thesis reports or final assignments, information from online news, and Indonesia's national Twitter.

3. *Text Pre-processing*: These processes help to make the text data more consistent and structured for analysis and to remove any irrelevant or redundant information that might impact the results. These activities include (i)case folding: converting all characters or capital letters to lowercase, (ii)tokenizing: breaking strings or sentences into individual words, and (iii)stop word removal: eliminating meaningless words from the list of tokens generated during tokenizing.

4. *Implementation of Processing Techniques*: The next stage is determining the technique or method used to carry out the text pre-processing process. The method used is Nazief Adriani and Tala.

5. *Testing and Comparing Results*: After the method has been determined, a stemming test is carried out using the two algorithms. Tests were carried out to see the results and compare these results both in terms of accuracy and speed.

6. *Evaluation of Results*: After testing and comparing the results, an evaluation is carried out. Evaluation relates to the accuracy and speed of the two methods used to carry out the stemming process.

7. *Conclusions*: The final stage is concluding the evaluation that has been carried out. This conclusion is related to comparing algorithms with high accuracy and speed in the stemming process.

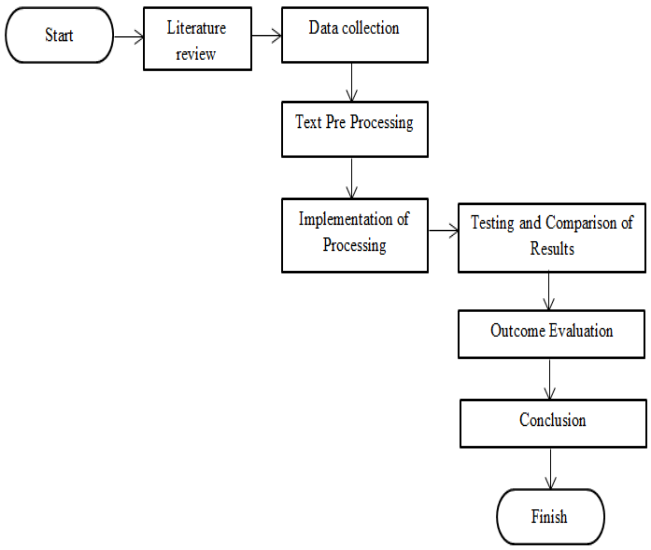

Figure 1. Research Method

B. Data Collection Methods

The data collection method is defined as a method used by scientists to obtain data to support their research activities. This method leads to the origin of the data to be obtained either. The main sources are interviews and observations or supporting sources in the form of literature studies through

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

books, journals, and proceedings [14]. Based on the understanding of data collection techniques, the data obtained in this activity comes from supporting sources, namely from several student thesis and final assignments, online news, Twitter, books, journals, and proceedings that discuss the Indonesian language text stemming process. The data obtained will be used as a reference to analyze the accuracy and speed of the stemming process from Nazief Adriani and Tala's algorithm.

C. Stemming

A stemming algorithm maps the different morphological variants of a set of words into basic words. This stemming algorithm is widely used in computational linguistics and information retrieval[15]. Stemming can also be interpreted as removing affixes, prefixes, suffixes, and prepositions. The result from the stopword removal process so that terms can become basic word forms [6], whereas according to [16], stemming is the process of finding basic words that are used to define features in the text. Figure 2 shows what exactly is meant by the stemming process.

Figure 2. An Example of Stemming

This study aims to analyze the performance of the Nazief Adriani algorithm with Tala for the Indonesian language text-stemming process. The expected result is to know the reliability of the two algorithms used in terms of accuracy and processing speed. Figure 3 explicitly shows the mindset of the stemming testing process. The data sources used in this study came from the online news site Tribunnews.com and extracts from student thesis reports.

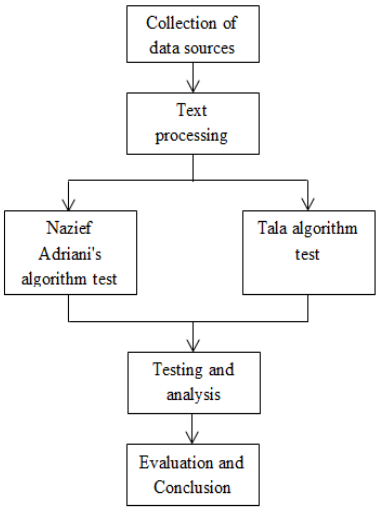

Figure 3. The Flow Of Thought Of The Stemming Process

D. Nazief Adriani's Algorithm

Nazief Adriani's algorithm is one of several stemming algorithms discovered by Bobby Nazief and Mirna Adriani. According to [2], several stages or steps for solving using the Nazief Adriani algorithm include:

1. Look for words that will be stemmed from the prepared dictionary. If these words are in the dictionary, it is concluded that the word is a root word, and the process stops.
2. Discarding Suffixes like "*-lah*", "*-kah*", "*-ku*", "*-mu*", or "*-nya*". If the suffix is a particle like "*-lah*",'*-kah*","*-tah*" or "*-pun*" then repeat the steps to remove possessive pronouns like "*-ku*", "*-mu*", or "*-nya*", when found.
3. Removing Derivation Suffixes ("-i", "-an", or "*-ka*n"). If the suffix is found in the dictionary, the process will stop, but if it is not found, the process will continue in step 3.1.
   3.1 If the suffix "-an" is deleted and the word ends in "-k", then the suffix is also deleted. If the word is found in the dictionary, the algorithm stops. If not found, then do step 3.2.
   3.2 Restores deleted suffixes such as "-i", "-an", or "*-kan*"
4. Remove Derivation Prefix. If in step 3 above there are deleted suffixes, then the process continues to step 4.1, but if no suffixes are deleted, then the process continues in step 4.2.
   4.1 Perform checks on Tables that contain prefix and suffix combinations that are not allowed to be used if it is found in the Table. The process stops; otherwise, if it is not found, it continues in step 4.2.
   4.2 Determine the prefix type and then delete the prefix. If you haven't found the root word, then the process continues in step 5, but conversely, if the root word is found, the process stops with a note that the position of the second prefix is the same as the first prefix.
5. Perform step recording
6. If all stages have been completed but have not produced results, the initial word can be concluded as the root word. Process complete.

The steps in determining the prefix type of the Nazief Adriani algorithm are as follows :

1. The prefix "*di-*", "*to-*", or "*se-*" has the prefix type "*di-*", "*ke-*", or "*se-*".
2. An additional process is needed to determine the type of the prefix if the prefix is known as "*te-*", "*me-*", "*be-*", or "*pe-*".
3. The process will stop if two first characters are not "*di-", "ke-", "se-", "te-", "be-", "me-", or "pe-*".
4. Type the prefix "none" then the process stops. However, if the type of the prefix is not "none", the determination of the prefix can be seen in Table II. Prefixes can be removed if found.

The following presents a list of combinations of prefixes and suffixes that are not permitted to be used, how to determine the type of prefix "te" and groups of prefixes based on the type of prefix as shown in Tables I, II, and III.

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

TABLE I
NOT PERMITTED COMBINATIONS OF PREFIGURE AND SUFFICIENCY

No	Prefix	Inappropriate ending
1	*be-*	-i
2	*di-*	*-an*
3	*ke-*	*-i,-kan*
4	*me-*	*-an*
5	*se-*	*-i, -kan*

TABLE II
DETERMINATION OF THE TYPE OF "*TE*"

Following Characters				Prefix Type
Set 1	Set 2	Set 3	Set 4	
-r-	"-r-"	-	-	none
-r-	Vowel	-	-	*ter-luluh*
-r-	not (vowel or "-r-")	"-er-"	vowel	*ter*
-r-	not (vowel or "-r-")	"-er-"	not vowel	*ter-*
-r-	not (vowel or "-r-")	not "-er-"	-	*ter*
not (vowel or "-r-")	"-er-"	Vowel	-	none
not (vowel or "-r-")	"-er-"	not vowel	-	*te*

TABLE III
GROUP OF THE PREFIX BY TYPE OF PREFIX

Prefix	Removed Prefix
di-	*di-*
ke-	*ke-*
se-	*se-*
te-	*te-*
ter-	*ter-*
ter-luluh	*Ter*

E. Tala Algorithm

The Tala algorithm performs processing from prefixes, suffixes, and combinations of prefixes and suffixes in derived words. Stemming Tuning is a steeming algorithm adopted from a special English stemming algorithm, namely Porter's stemming. The Indonesian Tala steamer has the following Indonesian word formation structure [17] :

[prefix-1] + [prefix-2] + base + [suffix] + [belongs to] + [carrying], where each of these parts is combined with a root word to form a word that has a reward. The Tala algorithm has three initial steps and two optional steps, as follows:

1. Perform particle removal
2. Eliminate Possessive Pronouns
3. The first prefix is removed. If no process is found, proceed to step 4.1, but if there is, do a search, and the process will continue at step 4.2
4. Stages step 4
   4.1. Remove the second prefix and go to step 5.1
   4.2. Remove suffix. If it is not found, the word you are looking for is concluded as a basic word, but if it is found, proceed to step 5.2.
5. Stages step 5
   5.1. Removing the ending and the final word is concluded as a base word.
   5.2. The second prefix is deleted, and the final word is concluded as a base word.

The Tala algorithm has five categories of affix rules shown in Table IV to Table VIII.

TABLE IV
RULES FOR THE INFLECTIONAL PARTICLE

Suffix	Replacement	Measure Condition	Additional Condition	Example
-kah	NULL	2	NULL	*bukukah*
-lah	NULL	2	NULL	*pergilah*
-pun	NULL	2	NULL	*bukupun*

TABLE V
RULES FOR THE INFLECTIONAL POSSESSIVE PRONOUN

Suffix	Replacement	Measure Condition	Additional Condition	Example
-ku	NULL	2	NULL	*bukuku*
-mu	NULL	2	NULL	*bukumu*
-nya	NULL	2	NULL	*bukunya*

TABLE VI
RULES FOR FIRST-ORDER DERIVATIONAL PREFIX

Suffix	Replacement	Measure Condition	Additional Condition	Example	
meng-	NULL	2	NULL	*mengukur*	*ukur*
meny-	S	2	V ... *	*menyapu*	*sapu*
men-	NULL	2	NULL	*menduga*	*duga*
mem-	P	2	V ...	*memaksa*	*paksa*
mem-	NULL	2	NULL	*membaca*	*baca*
me-	NULL	2	NULL	*merusak*	*rusak*
peng-	NULL	2	NULL	*pengukur*	*ukur*
peny-	S	2	V ...	*penyapu*	*sapu*
pen-	NULL	2	NULL	*penduga*	*duga*
pem-	P	2	V ...	*pemaksa*	*paksa*
pem-	NULL	2	NULL	*Pembaca*	*baca*
di-	NULL	2	NULL	*diukur*	*ukur*
ter-	NULL	2	NULL	*tersapu*	*sapu*
ke-	NULL	2	NULL	*kekasih*	*kasih*

TABLE VII
RULES FOR SECOND-ORDER DERIVATIONAL PREFIX

Suffix	Replacement	Measure Condition	Additional Condition	Example	
ber-	NULL	2	NULL	*berlari*	*lari*
bel-	NULL	2	Ajar	*belajar*	*ajar*
be-	NULL	2	k*er	*bekerja*	*kerja*
per-	NULL	2	NULL	*perjelas*	*jelas*
pel-	NULL	2	Ajar	*pelajar*	*ajar*
pe-	NULL	2	NULL	*pekerja*	*kerja*

TABLE VIII
RULES FOR DERIVATIONAL SUFFIX

Suffix	Repla-cement	Measure condition	Additional condition	Example	
-kan	null	2	prefix is not a member {*ke, peng*}	*tarikkan*	*tarik*
				mengambilkan	*ambil*
-an	null	2	prefix prefix is not a member {*di, meng, ter*}	*makanan*	*makan*
				perjanjian	*janji*
-i	null	2	prefix is not	*tandai*	*tanda*

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

Suffix	Repla-cement	Measure condition	Additional condition	Example	
			a member {*ber, ke, peng*}	*mendapati*	*dapat*

F. Text Pre-processing

In the text pre-processing stage, the first stage is cleaning of the data that has been collected, then the case folding, tokenizing, and stopword removal stages. The cleaning process is the process of removing characters, letters, and symbols that are outside the letters of the alphabet in text. Deleted letters or characters can be in the form of copyright, news endings, and copyright symbols. Then after cleaning, the next process is carried out, case folding. Case folding is the process of equating all letters to lowercase. The text processing flowchart is in Figure 4.

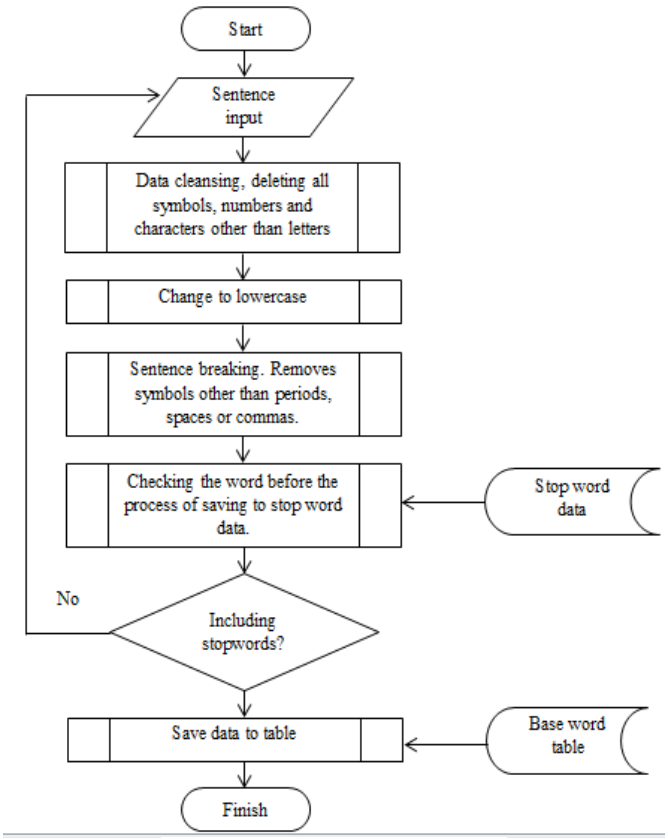

Figure 4. Pre-processing Text Flowchart

From the pre-processing text flow chart above, an explanation can be given as follows:

1. Enter A Sentence or Text: The text can come from the essence of the final project report or student thesis and news text obtained from the Tribunnews.com site

2. Cleaning process. The text or sentence entered will be checked to see whether there are other characters or symbols. If there is, then the character or symbol will be deleted.

3. Change to Lowercase: After cleaning and the text is considered clean. The capitalization of letters will be equalized. Namely, all letters will be converted to lowercase. This aims to facilitate the process of stemming and stopword removal.

4. Sentence Breaking: The process of solving sentences is commonly known as stemming, a process carried out to break sentences into words. This split is based on spaces, periods, or commas. Apart from these marks, other characters will be removed.

5. Word Check: This stage is to record whether the words resulting from the stemming process will be ignored in the Stopword removal process.

6. Save The Words: Ignored in the Stopword removal process into the base word Table

The stemming process can begin after completing the two stages described above. The process of stemming is used to separate sentences into individual words, and a process of checking will be carried out to determine whether or not stemming a word results in a word that will be ignored during the process of stopword removal. Stemming is a process that is used to separate sentences into individual words. The results of Tala's research were used as the stopword data in Table IX.

TABLE IX
STOPWORD DATA EXAMPLE (TALA 2013)

ada	*amatlah*	*atas*
adalah	*anda*	*or*
adanya	*andalah*	*orkah*
adapun	*antar*	*orpun*
agak	*antara*	*awal*
agaknya	*antaranya*	*awalnya*
agar	*apa*	*bagai*
akan	*apaan*	*bagaikan*
akankah	*apabila*	*bagaimana*
akhir	*apakah*	*bagaimanakah*
akhiri	*apalagi*	*bagaimanapun*
akhirnya	*apatah*	*bagi*
aku	*artinya*	*bagian*
akulah	*asal*	*...*
amat	*asalkan*	

A. *Stemming Process*

Stemming is a procedure that is a part of an information retrieval system (IR, which stands for "Information Retrieval"), and the purpose of this process is to transform the words in a text into their most fundamental forms by applying a set of rules. For instance, "same" is shown to be the root word of "bring," "bring," and "bring," when these words are subjected to stemming. According to the author of [16], stemming is the process of locating basic words, which are then employed in the subsequent step of determining features in the text. Figures 5 and 6 present the flow chart representation of the stemming process.

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

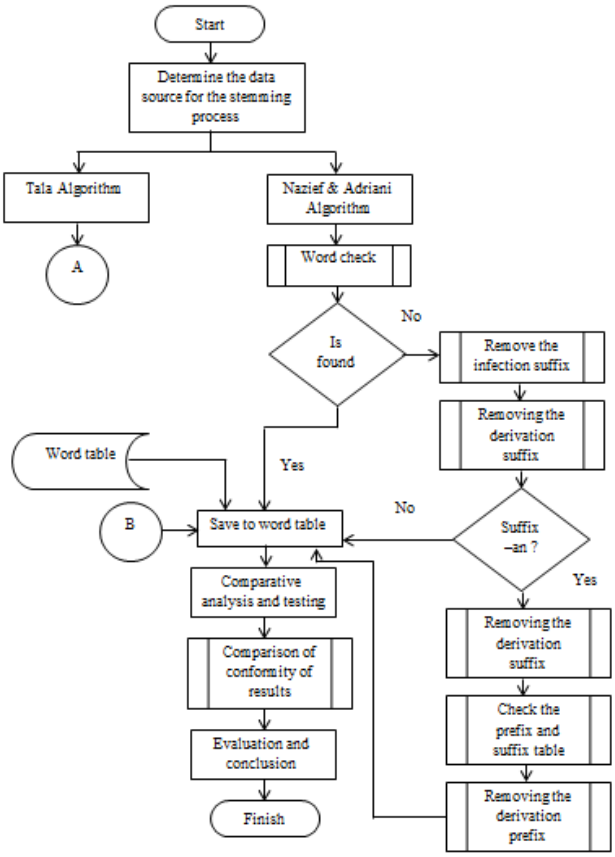

Figure 5. Flowchart Stemming 1

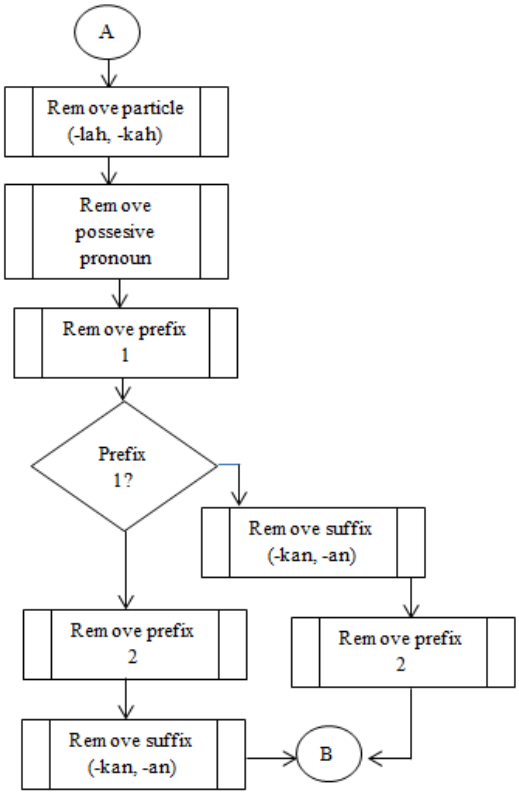

Figure 6. Flowchart Stemming 2

III. Result and Discussion

A. Test Data Collection

The data used for testing in this scenario is sourced from the news website Tribunnews.com and abstracts of student thesis documents. These data will be storedin tables X and XI database for further processing and analysis.

TABLE X
EXAMPLE OF NEWS IN DATABASE TABLE

NewsId	01
Nm_filenews	01.doc
URL	*https://www.tribunnews.com/nasional/2022/12/06/genap-45-tahun-bpjs-ketenagakerjaan-satukan-semangat-sejahterakan-pekerja*
Title_news	*Genap 45 Tahun, BPJS Ketenagakerjaan Satukan Semangat Sejahterakan Pekerja*
News content	*Genap memasuki usia 45 tahun, BPJS Ketenagakerjaan berikrar ùntuk terus berkembang dan bergerak maju, menjaga integritas serta menyatukan semangat mensejahterakan seluruh pekerja Indonesia. Anggoro Eko Cahyo dalam keterangannya mengatakan, pihaknya berkomitmen menyatukan semangat yang datang dari dari seluruh insan BPJS Ketenagakerjaan dan juga dari stakeholders terdekat seperti kementerian, pengusaha hingga serikat pekerja/ buruh, untuk meningkatkan kesejahteraan pekerja Indonesia. "Hari ini kami genap berusia 45 tahun, sebuah usia yang sudah bisa dibilang matang, kami berikrar untuk terus memperluas cakupan perlindungan program jaminan sosial ketenagakerjaan untuk seluruh pekerja, terutama saat ini untuk pekerja informal or pekerja bukan penerima upah, dan juga kami akan terus meningkatkan kualitas pelayanan, sehingga peserta akan semakin merasakan manfaat hadirnya BPJS Ketenagakerjaan," ucap Anggoro. Diketahui, yang dilakukan BPJS Ketenagakerjaan saat ini adalah mengoptimalkan strategi ekstensifikasi, intensifikasi dan retensi. Memanfaatkan peluang kerja sama dengan Kementerian/Lembaga dan Pemerintah Daerah, business to business, serta utilisasi engine PERISAI, dan dikarenakan target peserta adalah BPU, kampanye Kerja Keras Bebas Cemas akan digunakan untuk melindungi sebanyak-banyaknya pekerja. "Saat ini pencapaian kepesertaan aktif kami adalah sebesar 36 juta tenaga kerja or meningkat 6 juta dari tahun sebelumnya. Angka peningkatan ini merupakan rekor tertinggi selama BPJS Ketenagakerjaan berdiri, dan target sampai dengan tahun 2026 adalah 70 juta tenaga kerja," ungkapnya. Selama tahun 2022, kinerja pelayanan BPJS Ketenagakerjaan juga terus meningkat. Komitmen perubahan mindset ke arah customer oriented telah membawa perubahan terhadap kualitas manfaat dan layanan yang terasa makin dekat dengan peserta. Tercatat success rate Jaminan Hari Tua (JHT) tahun ini telah mencapai 99.58 persen, dengan rata-rata SLA masa tunggu JHT via Online or video call kurang dari 3 hari, serta rata-rata proses klaim JHT via Jamsostek Mobile (JMO) kurang dari 15 menit. Utilisasi kanal klaim melalui aplikasi JMO juga tercatat di angka 25%, lebih tinggi dari kanal Kantor Cabang sebesar 15%, namun masih dibawah utilisasi kanal Online (video call) sebesar 60%.*

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

TABLE XI
EXAMPLE OF THE THESIS DOCUMENT ABSTRACT

Kddok	D01
Nm_dok	Faiz-skripsi.doc
Abstract Contents	Penyalaan pada sepeda motor yang sekaligus juga berfungsi sebagai pengamanan sepeda motor harus dirancang dan dibuat seaman mungkin untuk menghindari hilangnya kendaraan. Peralatan yang dirancang ini untuk menghidupkan dan mematikan sepeda motor dengan sistem operasi android melalui jaringan Hotspot/Wi-Fi. Input pada sistem operasi android dilakukan pada web browser dan asisten google, lalu data diteruskan dan dikontrol menggunakan mikrokontroler ESP32. Dari hasil pengujian diperoleh bahwa jarak maksimum yang dapat dicapai Wi-Fi antara android dengan mikrokontroler ESP32 yang berada pada sepeda motor untuk mengoperasikan engine sepeda motor adalah kurang-lebih sekitar 16 meter. Sistem ini juga membuat penyalaan engine sepeda motor menjadi penyalaan pintar karena dalam penyalaan sepeda motor pengemudi dapat menggunakan perintah suara menggunakan asisten google. Pengamanan sistem ini tidak terbatas pada pengamanan saat sepeda motor stationary or pada saat sepeda motor berada diparkiran, pengamanan ini juga bisa berada pada saat kendaraan digunakan, dan ketika sepeda motor ditinggalkan pengemudi pada saat menyala, maka sepeda motor akan otomatis mati jika mikrokontroler ESP32 berada diluar jangkauan Wi-Fi/Hotspot yang berasal dari perangkat android. Kata Kunci : Sepeda Motor, Android, Mikrokontroler ESP32, Asisten Google

B. Accuracy And Speed Test

Determine the degree of precision that is required. It is possible to accomplish this by counting the words found and then dividing those results by the total number of words contained in a single dataset. A comparative display of the test outcomes using the two approaches is created to determine the degree of precision achieved by each processing step following the stemming process. The information to be gathered will be presented in the following formats: total words, correct words, incorrect words, accuracy, and time [2].

A large amount of test data is needed to obtain correct accuracy results. Table XII shows an example of text testing results that have been stemmed using one of the two methods that will be tested in this study.

TABLE XI
EXAMPLE OF TEST RESULTS WITH ONE TECHNIQUE

No	Code	Total	True	False	True %	False %	Time
1	012	254	180	68	70,8	26,77	65,70
2	018	110	85	30	77,2	27,27	20,99
3	019	170	155	15	91,1	8,82	40,00
4	021	127	108	15	85,0	11,81	20,20
5	023	150	90	68	60	45,33	40,14
...	...	...	...	...	...	...	...
500	026	70	63	4	90	5,714	8,054
	Total	58,959	46,4	12,5	39, 6	12,02	58,95
	Average	115,430	91,7	24,7	78,5	21,53	23,75

The data in Table XII is an example of the test results with the news data source *tribunnews.com* which has 500 news stories with a total of 58,959 words. The average obtained is around 116 words in the news dataset with an average accuracy of 78.46% for the news dataset.

C. Analysis of Test Results

In this study, the source data used for the accuracy testing process using the Nazief Andriani and Tala method came from news data on the Tribunnews.com website and student thesis report abstracts. The variable for measuring the accuracy value is based on previous research by [2]. The measurement pattern is obtained from the number of words found compared to the total words in the dataset. The testing process was carried out using a total of 4 scenarios for each stemming method used. Testing is done to get the value of accuracy and speed. The test scenario data is in Table XIII.

TABLE XII
TEST SCENARIO DATA

Scenario	Data Source	Document	Algorithm
X1	News text from the Tribunnews.com site	506	Nazief & Adriani
X2	Teks berita dari situs Tribunnews.com	506	Tala
Y1	The text of the thesis report abstract	50	Nazief & Adriani
Y2	The text of the thesis report abstract	50	Tala

Based on the data in Table XII, the first test scenario will be carried out, namely testing data originating from the news site Tribunnews.com where a total of 500 news documents are obtained. The stemming algorithm used for scenario X1 is the Nazief & Andriani method, and scenario X2 uses the Tala method. . The first stage is to carry out the text pre-processing process and continue with the stemming process. After testing the first scenario, proceed with testing the second scenario, Y1, and Y2, where the data source comes from the abstract of student thesis reports. The Y1 test uses the Nazief & Andriani algorithm, while the Y2 test uses the Tala algorithm.

From the scenario that is carried out next, accuracy measurements will be carried out. Accuracy measurement is carried out based on the total number of words found to have a basic word type divided by the total number of words in one dataset. In addition, the calculation of how much time is needed by the method in the stemming process with the data sources provided. Time calculation is calculated from the difference in time after the process is complete minus the time the process starts. From these results, it will be obtained the average of each scenario for accuracy and processing time. After carrying out a series of stemming tests with the two algorithms used, the results are obtained as shown in Tables XIII and XIV.

TABLE XIII
TEST RESULT DATA X1

No	Code	Total	True	False	True %	False %	Time
1	012	254	180	68	70,8	26,77	65,70
2	018	110	85	30	77,2	27,27	20,99
3	019	170	155	15	91,1	8,82	40,00
4	021	127	108	15	85,0	11,81	20,20
5	023	150	90	68	60	45,33	40,14
...	...	...	...	...	...	...	...

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

No	Code	Total	True	False	True %	False %	Time
500	026	70	63	4	90	5,714	8,054
	Total	58,959	46,4	12,5	39, 6	12,02	58,95
	Average	115,430	91,7	24,7	78,5	21,53	23,75

TABLE XIV
TEST RESULT DATA X2

No	Code	Total	True	False	True %	False %
1	300712	254	180	68	70,86	65,70
2	300718	110	85	30	77,27	20,993
3	300719	170	155	15	91,17	40,00
4	300721	127	108	15	85,03	20,20
5	300723	150	90	68	60	40,14
...	...	...	...	...	...	...
500	300726	70	63	4	90	8,054
	Total	59,652	34,232	20,331		5,022
	Average	117,230	86,56	35,24	64,37	10,75

From the data from the trial process shown in Table XIII and Table XIV, it can be concluded that for test scenarios 1 X1 and X2 with data originating from news on the Tribunnews.com site, the total words in Table XIII are 59,959, with an average of 115 say. The average accuracy value obtained was 78.47% and the average time needed to process was 23.75 seconds, while for Table XIV the total words obtained were 59,652 with an average of 117 words. For accuracy obtained with an average value of 64.37% and an average time required of 10.75 seconds. Based on these data, the results of testing the accuracy of the news source data Tribunnews.com, the Nazief & Adriani algorithm has a higher value than Tala, which is 78.47%. In contrast, in terms of processing time, the Tala algorithm is faster than using Nazief & Adriani, which is 10.75 seconds. After testing the X1 and X2 scenarios, the process of testing the Y1 and Y2 scenarios was then carried out, where the data for this test came from the abstract of the student thesis document. From the tests carried out, the results are in Table XV and Table XVI.

TABLE XV
TEST RESULT DATA Y1

No	File Name	Total	True	False	True %	False %
1	a.pdf	374	310	64	82,89	226, 7
2	b.pdf	432	402	30	93,06	308,8
3	c.pdf	156	140	16	89,74	68,0
4	d.pdf	175	153	22	87,43	75,6
5	e.pdf	142	122	20	85,92	104,1
...	..	…	…	…	…	…
50	n.pdf	117	101	16	86,32	44,384
Total		7.438	6.114	1.211		3.407,9
Average		148	124,68	26,28	82,63	65,2

TABLE XVI
TEST RESULT DATA Y2

No	File Name	Total	True	False	True %	False %
1	a.pdf	190	144	46	75,8	53,6
2	b.pdf	238	179	59	75,2	6,8
3	c.pdf	113	90	23	79,6	15,9
4	d.pdf	178	131	47	73,6	28,5
5	e.pdf	142	102	40	71,8	33,3
...	…	…	…	…	…	…
50	n.pdf	127	90	37	70,9	27,9
	Total	7.067	5.013	2.053		1.705
	Average	140	99,420	40,100	68,6	32,19

From the test results presented in Tables XV and XVI, it can be explained that Table XV used the Nazief & Andriani algorithm as the stemming method and found 7,438 words, with an average of 148 words. The average value of accuracy was obtained at 82.63%, with an average time required of 65.2 seconds, while for Table XVI using the Tala algorithm, 7,067 words were found with an average of 140 words. An average value of 68.6% is obtained for accuracy with an average time of 34.19 seconds. Based on the XV and XVI Table data above, it can be concluded that the accuracy algorithm from Nazief & Andriani has a higher value than Tala, which is 82.63%. In contrast, regarding the time required, the Tala algorithm has a faster time than the Nazief & Andriani algorithm. Adriani. ie 32.19 seconds.

IV. CONCLUSION

The average accuracy value obtained using a new dataset of 500 news for the Nazief & Andriani method is 78.47%, totaling 58,964 words. The Tala method's average accuracy is 65.29%, with a total of 59,652 words. The average accuracy value obtained using a document dataset of 50 thesis reports for the Nazief & Andriani method is 82.63% % with a total of 7,438 words. In contrast, the Tala method has an accuracy value of 68.6%, totaling 7,067 words. The average value of the fastest processing time using the new dataset for the Nazief & Andriani method is 25.27 seconds, and the Tala method is 11.08 seconds. For the document dataset, the average processing time is the fastest, with the Nazief & Adriani method of 65.2 seconds and the Tala method of 32.19 seconds.

In terms of the accuracy of the stemming process, the Nazief & Andriani algorithm has better accuracy than the Tala algorithm. On the other hand, in terms of processing time speed, the Tala algorithm has better speed than the Nazief & Andriani algorithm.

Further work is being developed to better understand the stemming algorithm's reliability in terms of the accuracy and speed of the stemming process. The dataset needs to be reproduced not only from two sources but from at least three data sources. Many data sources are needed so that the test results can be more optimal.

REFERENCES

[1] Afuan L, "Stemming Indonesian Text Documents Using Porter's Algorithm". *Telematics Journal* Vol. 6 No. 2, pp. 34-40, 2013.

[2] Agusta L. "Comparison of Porter's Stemming Algorithm with Nazief & Adriani's Algorithm for Stemming Indonesian Text Documents". *System and Informatics National Conference 2009*, November 2009.

[3] Novitasari, D., "Comparison Of Porter's Stemming Algorithm With Arifin Setiono To Determine The Level Of Accuracy Of Basic Word", *String Journal*, Vol. 1 No. 2, pp. 120-129, 2016.

[4] Nopiyanti D., Sekarwati, K.A, "Basic Word Search Application for Indonesian Language Documents Using the Porter Stemming Method Using PHP & MYSQL", *Proceedings of the National Scientific*

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569

Seminar on Computers and Intelligence Systems (KOMMIT 2014). Oktober 2014.

[5] Utomo, M.S., "Tala Stemmer Implementation in Web-Based Applications". *DYNAMIC Information Technology Journa,l* Vol. 18, No. 1. Pp. 41-45, ISSN : 0854-9524, 2013.

[6] Wiguna, P.B,S., Hantono,B.S., "Improvement of the Indonesian Porter Stemmer Algorithm based on the Morphological Method by Applying 2 Morphological Levels and Prefix and Suffix Combination Rules",*JNTETI,* Vol. 2 No. 2, pp. 1-6, ISSN : 2301 – 4156, 2013.

[7] Indriyono, B.V, Utami E, Sunyoto, A. "Utilization of the Porter Stemmer Algorithm for Deep Indonesian Book Type Classification Process", *Journal of Informatics Buana,* Vol. 6 No.4, pp. 301-310, 2015.

[8] Ariyani, P.F, , Rahmala A., Juliasari, N." Implementation of Tala Stemming Method and Jaccard Function In the Library Catalog Application", *National Seminar on Innovation and Technology Application in Industry 2019*, February 2019.

[9] Ghazvini, A., et al, "Stemming algorithm for different tenses to improve Persian dictionary", 2012 *IEEE Symposium on Industrial Electronics and Applications*, September 2012.

[10] Pramudita, H.R., "Implementation Of Nazief & Adriani's Stemming Algorithm And Similarity On Acceptance Of Thesis Title", *DASI Scientific Journal*, Vol. 15 No. 04, pp. 15-19, ISSN : 1411-3201

[11] Yulianto, M.A.,, Nurhasanah, "The Effect of Stemming Nazief & Adriani on the Performance of the Rabin-Karp Algorithm in Detecting Text Similarities", *Pamulang University Informatics Journal,* Vol. 6, No. 4, pp. 880-886, ISSN : 2541-1004, 2021.

[12] Sugiyono, *Quantitative Research Methods, Qualitative, and R&D.*Bandung : Alfabeta. 2017

[13] Hasibuan, Z.. Research Methodology in the Field of Computers and Information Technology. Jakarta : University of Indonesia.

[14] Arikunto, S. *Research procedure*. Jakarta: Rineka cipta.

[15] Parwita, W.G.S, "Testing the Accuracy of Content-Based Filtering Recommendation Systems", *Mulawarman Informatics: Scientific Journal of Computer Science,* Vol. 14, No. 1, pp. 27-32, ISSN : 1858-4853, 2019.

[16] Prihatini, P.M, et al, "Stemming Algorithm for Indonesian Digital News Text Processing", *International Journal of Engineering and Emerging Technology*, Vol. 2, No. 2, pp. 1-7, ISSN : 2579-5988, 2017.

[17] Saifudin, A., Verdaningroem, N.J,M.," Application of the Basic Dictionary on Porter's Algorithm to Reduce Indonesian Stemming Errors", *Technology Journal,* Vol. 10, No.2 , pp. 103-112. ISSN : 2085 – 1669, 2018.

DOI : http://dx.doi.org/10.25139/inform.v8i1.5569